# Using Machine Learning to Detect Cognitive States across Multiple Subjects

**Xuerui Wang, Tom M. Mitchell and Rebecca Hutchinson**
Center for Automated Learning and Discovery
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
{xuerui, tom.mitchell, rah}@cs.cmu.edu

## Abstract

Is it feasible to train cross-subject classifiers to decode the cognitive states of human subjects based on functional Magnetic Resonance Imaging (fMRI) data observed over a single time interval? If so, these trained classifiers could be used as virtual sensors to detect cognitive states that apply across multiple human subjects. This problem is relevant to experimental research in cognitive science and to diagnosis of mental processes in patients with brain injuries. The biggest obstacle to training inter-subject classifiers on fMRI data is anatomical variability among subjects. We describe two approaches to overcoming this difficulty. The first approach takes advantage of the anatomically defined Region of Interest (ROI) as a basis for spatially abstracting the data, and the second one transforms the data from different subjects into Talairach-Tournoux coordinates. In particular, we present two case studies in which we have successfully trained cross-subject classifier to distinguish cognitive states such as (1) whether the human subject is looking at a picture or a sentence describing that picture, and (2) whether the subject is reading an ambiguous or unambiguous sentence.

## 1 Introduction

The study of human brain function has received a tremendous boost in recent years from the advent of fMRI, a brain imaging method that dramatically improves our ability to observe correlates of neural brain activity in human subjects at high spatial resolution (several millimeters), across the entire brain. This fMRI technology offers the promise of revolutionary new approaches to studying human cognitive processes, provided we can develop appropriate data analysis methods to make sense of this huge volume of data. A typical twenty-minute fMRI session with a single human subject produces a series of three dimensional brain images each containing approximately 15,000 voxels, collected once per second, yielding tens of millions of data observations.

*Corresponding author. http://www.cs.cmu.edu/∼xuerui; Tel: +1-412-268-1294; Fax: +1-412-268-3431

Since its advent, fMRI has been used to conduct hundreds of studies that identify specific regions of the brain that are activated on average when a human performs a particular cognitive function (e.g., reading, mental imagery). The vast majority of this published work reports descriptive statistics of brain activity, calculated by *averaging together* fMRI data collected over multiple time intervals, in which the subject responds to repeated stimuli of some type (e.g., reading a variety of words).

In our previous studies [7,8,9], we have successfully trained machine learning classifiers to automatically decode the cognitive state of single human subjects, given just their fMRI activity at a single time instant or time interval. We would like to extend those classifiers to multiple subjects. This goal of training cross-subject classifiers to detect cognitive states is important because such classifiers could provide the basis for new approaches to studying human reasoning processes in both normal and abnormal populations. Put succinctly, such classifiers would constitute *virtual sensors* of human subjects' cognitive states, which could be useful to scientists and clinicians across a range of cognitive science research and diagnostic medical applications.

This paper is organized as follows. We first provide a brief introduction to fMRI in Section 2, then present an overview of related work in Section 3, and the machine learning methods we used in this paper are enumerated in Section 4. The two fMRI data sets we analyze are described in Section 5 in which our previous results for single subject classifiers are also summarized for comparison purpose. In Section 6, we describe and compare the two methods we use to train cross-subject classifiers, and present our results.

## 2 Functional Magnetic Resonance Imaging

Functional Magnetic Resonance Imaging (fMRI) is a technique for obtaining three-dimensional images related to activity in the brain through time. More precisely, fMRI measures the ratio of oxygenated hemoglobin to deoxygenated hemoglobin in the blood with respect to a control baseline, at many individual locations within the brain. It is widely believed that blood oxygen level is influenced by local neural activity, and hence this blood oxygen level dependent (BOLD) response is generally taken as an indicator of neural activity.

An fMRI scanner measures the value of the fMRI signal (BOLD response) at all the points in a three dimensional grid (or *image*), covering part of the brain. In the two studies described in this paper, a three dimensional image is captured every 0.5 or 1.5 seconds. We refer to the cells within an image as *voxels* (volume elements). The voxels in a typical fMRI study have a volume of a few tens of cubic millimeters, and a three dimensional image typically contains tens of thousands of voxels, 10,000 to 15,000 of which contain cortical matter and are thus of interest. While the spatial resolution of fMRI is dramatically better than that provided by earlier brain imaging methods, each voxel nevertheless contains on the order of hundreds of thousands of neurons.

The temporal response of the fMRI BOLD signal is smeared over several seconds. Given an impulse stimulus such as a flash of patterned light, the fMRI BOLD response increases to a maximum after approximately four to five seconds, typically returning to baseline levels after another five to ten seconds. To illustrate this, a small portion of fMRI data is illustrated in Figure 1.

## 3 Related Work

While there has been little work on our specific problem of training classifiers to decode cognitive states across multiple subjects, there are several papers describing work with

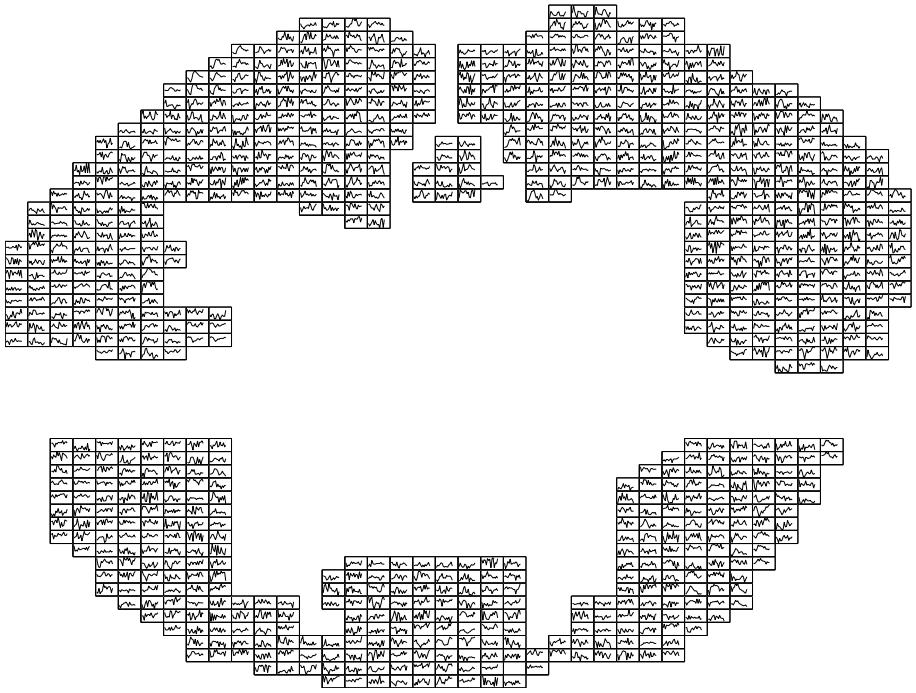


Figure 1: Typical fMRI data for a selected set of voxels in the cortex, from a two-dimensional image plane through the brain.

closely related goals. For example, Haxby et al. showed that different patterns of fMRI activity are generated when a human subject views a photograph of a face versus a house, versus a shoe, versus a chair [3]. While they did not specifically use these discovered patterns to classify subsequent single-event data, they did report that by dividing the fMRI data for each photograph category into two samples, they could automatically match the data samples related to the same category. Others (Wagner et al.) reported that they have been able to predict whether a verbal experience will be remembered later, based on the magnitude of activity within certain parts of left prefrontal and temporal cortices during that experience [11].

Our previous studies [7, 8, 9] demonstrated the feasibility of training classifiers to discriminate cognitive states of a human subject in several case studies. We also extended our classifiers across multiple subjects and across different contexts. This paper not only covers the conclusions in [9], but also includes our latest experimental results about cross-subject classifiers.

In addition to work on fMRI, there has been related recent work applying machine learning methods to data from other devices measuring brain activity. For example, Blankertz et al. described experiments training classifiers for single trial EEG data [1].

## 4 Approaches

This section briefly describes our approach to training classifiers, evaluating them, and selecting features. In our experiments all voxel activity values were represented by the percent difference from their mean value during fixation (rest) conditions.

### 4.1 Learning Method

In this paper we explore the use of machine learning methods to approximate classification functions of the following form

$$f : \text{fMRI-sequence}[I_1, ..., I_n] \rightarrow \text{CognitiveState}$$

where fMRI-sequence$[I_1, ..., I_n]$ is the sequence of fMRI images from $I_1$ to $I_n$ collected during a contiguous time interval and where CognitiveState is the set of cognitive states to be discriminated.

We explored a number of classifier training methods, including:

- *Gaussian Naive Bayes (GNB).* We assume each feature is independent to each other, and Gaussian distributed given the class label (see, for instance, [6]).
- *Support Vector Machine (SVM).* We use a linear kernel Support Vector Machine (see, for instance, [2]).
- *k Nearest Neighbor(kNN).* We use $k$ Nearest Neighbor with a Euclidean distance metric, considering values of 1, 3, and 5 for $k$ (see, for instance, [6]).

### 4.2 Results Evaluation

Trained classifiers are evaluated by their cross-validated classification accuracy when learning Boolean-valued classification functions. More precisely, we generally employ $k$-fold cross-validation, where $k$ is equal to the number of subjects in the corresponding study. In the following, we will call it *Leave one subject out cross validation*. In particular, for each subject we trained a classifier on the other $k-1$ subjects, measured the accuracy on the held out subject, and then calculated the mean accuracy over all held out subjects. Note that in our evaluation procedure, the competing classes are always balanced since we have equal numbers of examples of different classes in each subject. Our previous study [9] showed that imbalance between classes will greatly influence the performance of classifiers in the fMRI realm where the data are high dimensional and sparse.

### 4.3 Feature Selection

We explored a variety of methods for encoding an fMRI-sequence$[I_1, ..., I_n]$ as input to the classifier. In some cases, we encoded it as a vector of features, one for each voxel at each time in this interval. This can be an extremely high dimensional feature vector, consisting of hundreds of thousands of features given that a typical image contains 10,000 to 15,000 voxels, and a training example can include dozens of images. Therefore, it is natural to consider feature selection methods to reduce the dimensionality of the data before training the classifier. We explored a variety of approaches to reducing the dimension of this feature vector, including methods for feature selection, as well as methods that replace multiple feature values by their mean which help reduce the huge noise among fMRI data. Two kinds of feature selection methods were studied, discriminability methods and activity-based methods. The first one greedily selects the voxels having highest mutual information with the class labels, and the intuition behind the second one is that it emphasizes choosing voxels with large signal-to-noise ratios, though it ignores whether the feature distinguishes the target classes. Surprisingly, activity-based methods outperformed discriminability-based methods in most situations [9]. This paper will not discuss discriminability-based methods. Those feature selection and feature abstraction methods related to cross-subject classifiers are described as follows:

- *Average.* We average all voxels in an ROI into a supervoxel.

- *ActiveAvg(n).* For each ROI, we first select the $n$ most active voxels, and then average the selected voxels into a supervoxel.
- *Active(n).* We select the $n$ most active voxels in the whole brain.

For the task of training multiple subject classifiers, the first two will be used in the ROI mapping method since they are related to ROIs, and the last one will be used for the Talairach transformation method. ROI mapping and Talairach transformation will be discussed in detail in Section 6. In [9] we also designed many other activity-based methods. For instance, *RoiActive(n)* selects the $n$ most active voxels in each ROI. This kind of methods were shown useful in training single subject classifiers, but currently there is no convincing evidence that they could be used across subjects.

## 5 Case Studies

This section describes two case studies, as well as the results on single subject classifiers in these studies (detailed in [9]). We will give the results about the multiple subject classifiers in these studies in Section 6.

### 5.1 Sentence versus Picture Study

In this fMRI study [4], subjects performed a sequence of trials, during which they were first shown a sentence and a simple picture, then asked whether the sentence correctly described the picture. We used this data to explore the feasibility of training classifiers to distinguish whether the subject is examining a sentence or a picture during a particular time interval.

In half of the trials the picture was presented first, followed by the sentence, which we will refer to as *SP data set*. In the remaining trials, the sentence was presented first, followed by the picture, which we will call *PS data set*. In either case, the first stimulus (sentence or picture) was presented for 4 seconds, followed by a blank screen for 4 seconds. The second stimulus was then presented for up to 4 seconds, ending when the subject pressed the mouse button to indicate whether the sentence correctly described the picture. Finally, a rest or fixation period of 15 seconds was inserted before the next trial began. Thus, each trial lasted approximately 27 seconds. Pictures were geometric arrangements of the symbols $+$, $*$ and/or \$, such as

$$\begin{array}{c} + \\ \hline * \end{array}$$

Sentences were descriptions such as "It is true that the star is below the plus." Half of the sentences were negated (e.g., "It is not true that the star is above the plus.") and the other half were affirmative sentences.

Thirteen subjects participated in this study, and each subject was presented a total of 40 trials as described above, interspersed with ten fixation periods. In each fixation period the subject simply stared at a fixed point on the screen. fMRI images were collected every 0.5 seconds.

The learning task we consider for this study is to train a classifier to determine, given a particular 16-image interval of fMRI data, whether the subject was viewing a sentence or a picture during this interval. In other words, we wish to learn a classifier of the following form

$$f : \text{fMRI-sequence}[I_1, ..., I_{16}] \rightarrow \{\text{Picture, Sentence}\}$$

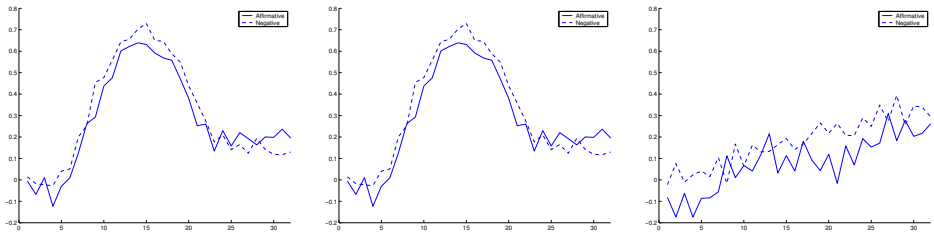
where $I_1$ is the time of stimulus (picture or sentence) onset. In this case 7 ROIs selected to be most relevant by a domain expert are used in ROI mapping method. These 7 ROIs

contained a total of 1397 to 2864 voxels per subject, varying due to differences in brain structure from one subject to another.

In this study, note that we extract one sentence example and one picture example from each trial, which is different from the Syntactic Ambiguity study where we get one example from each trial (see Section 5.2). We found that the intensity of most voxels in affirmative trials is lower than in negated trials although the response pattern is rather similar. The difference is shown in Figure 2. It is natural to consider some procedure to make the data from different trials become more similar. We employ the following normalization procedure[1]:

$$Y_t = \frac{X_t - min_t X_t}{max_t X_t - min_t X_t}$$

where $X_t$'s and $Y_t$'s are the data before and after normalization, respectively. This method linearly re-scales the data of each voxel in each trial into [0, 1]. From Table 1, we can find this simple procedure improves the accuracies of single subject classifiers in most cases.




(a) ROI A (b) ROI B (c)ROI C

Figure 2: Intensity difference between affirmative trials and negated trials. The activity value above is the mean activity of all subject in SP data set. The plots actually show the activity of supervoxels got by averaging all voxels in the corresponding ROIs.

The expected classification accuracy from random guessing in this case is 50%, given the equal number of examples from both classes. Some average accuracies of single subject classifiers in this study are summarized in Table 1.

### 5.2 Syntactic Ambiguity Study

In this fMRI study [5], subjects were presented with two types of ambiguous sentences and two types of unambiguous sentences, and were asked to respond to a yes-no question about the content of each sentence. Five normal subjects participated in this study. The questions were designed to ensure that the subject was in fact processing the sentence. The learning task for this study was to distinguish whether the subject was currently reading the least ambiguous or the most ambiguous type of sentence. An example of the most ambiguous type of sentence is "The experienced soldiers warned about the dangers conducted the midnight raid." An example of the least ambiguous type of sentence is "The experienced soldiers spoke about the dangers before the midnight raid."

Ten sentences of each of type were presented to each subject. Each sentence was presented for 10 seconds. Next a question was presented, and the subject was given 4 seconds to answer. After the subject answered the question, or 4 seconds elapsed, an "X" appeared on the screen for a 12 second rest period. The scanner collected one image every 1.5 seconds.

[1] We tried some more complex procedures in order to avoid the influence of outliers. For example, we use the 5% and 95% percentiles instead of maximum and minimum. The complex procedures did not perform better.

Table 1: Average accuracies of single subject classifiers in the Sentence versus Picture study. The accuracies within parenthesis are with normalization.

| METHOD | CLASSIFIER | SP | PS | SP+PS |
|---|---|---|---|---|
| Average | GNB | 86.5% (90.6%) | 72.5% (79.6%) | 69.6% (66.5%) |
| Average | SVM | 87.7% (89.0%) | 76.5% (83.7%) | 69.2% (69.8%) |
| Average | 1NN | 82.1% (86.5%) | 60.6% (61.9%) | 62.8% (59.7%) |
| Average | 3NN | 85.2% (87.5%) | 64.8% (69.2%) | 64.6% (59.7%) |
| Average | 5NN | 84.0% (89.4%) | 66.9% (74.6%) | 65.9% (60.4%) |
| ActiveAvg(20) | GNB | 89.0% (95.4%) | 76.0% (88.1%) | 72.1% (75.4%) |
| ActiveAvg(20) | 1NN | 88.8% (94.4%) | 71.2% (82.5%) | 72.0% (71.2%) |
| ActiveAvg(20) | 3NN | 89.8% (95.4%) | 75.4% (83.7%) | 76.6% (73.2%) |
| ActiveAvg(20) | 5NN | 90.0% (95.0%) | 76.7% (86.2%) | 76.6% (73.2%) |
| Active(140) | GNB | 91.5% (96.9%) | 80.0% (89.0%) | 79.2% (84.3%) |
| Active(140) | 1NN | 87.3% (94.0%) | 71.7% (83.5%) | 79.2% (86.0%) |
| Active(140) | 3NN | 89.2% (96.0%) | 76.9% (86.5%) | 80.8% (87.0%) |
| Active(140) | 5NN | 87.9% (96.3%) | 75.6% (86.9%) | 79.1% (86.1%) |

Table 2: Average accuracies of single subject classifiers in the Syntactic Ambiguity study

| METHOD | CLASSIFIER | AVERAGE ACCURACY |
|---|---|---|
| Average | GNB | 61% |
| Average | SVM | 63% |
| Average | 1NN | 54% |
| Average | 3NN | 64% |
| Average | 5NN | 60% |
| ActiveAvg(20) | GNB | 66% |
| ActiveAvg(20) | 1NN | 55% |
| ActiveAvg(20) | 3NN | 56% |
| ActiveAvg(20) | 5NN | 57% |
| Active(80) | GNB | 69% |
| Active(80) | 1NN | 57% |
| Active(80) | 1NN | 60% |
| Active(80) | 5NN | 61% |

We are interested here in learning a classifier that takes as input an interval of fMRI activity, and determines which of the two types of sentence the subject is reading. We trained classifiers of the form

$$f : \text{fMRI-sequence}[I_1, ..., I_{16}] \rightarrow \{\text{Ambiguous, Unambiguous}\}$$

where $I_1$ is the image at which the sentence is first presented to the subject. In this case 4 ROIs considered to be most relevant by a domain expert are used in ROI mapping method. These 4 ROIs contained a total of 1500 to 3508 voxels, depending on the subject.

Note that in this study, we actually want to distinguish trials, so the normalization method in previous section will have no effect. Given the equal number of ambiguous and unambiguous trials, the expected accuracy from random guessing is again 50%. Some average accuracies of single subject classifiers in this study are summarized in Table 2.

## 6 Experimental Results

Returning to our main question whether it is possible to train classifiers that apply across multiple human subjects, including subjects beyond the training set, it is easy to see that the biggest obstacle to inter-subject analysis of fMRI data is anatomical variability among subjects. Different brains have different shapes and sizes, making it problematic to register the many thousands of voxels in one brain to their precise corresponding locations in a second brain. One common approach in neuroscience to this problem is to transform (geometrically morph) fMRI data from different subjects into some standard anatomical space, such as Talairach coordinates [10]. However, some feature selection and abstraction methods used in our studies are already immune to anatomical variability. For example, by averaging the voxels in a particular ROI into a supervoxel (and treating it as a single voxel ROI afterwards), we can easily map one brain to another in terms of these anatomically defined ROIs. Both methods are used to combine data from different subjects in our studies. ROI mapping takes advantage of the anatomically defined ROI as a basis for spatially abstracting the data, and Talairach transformation converts the data from different subjects into the standard Talairach-Tournoux coordinates. Table 3 compares the upside and downside of the two methods.

Table 3: ROI mapping versus Talairach transformation

| | **ROI MAPPING** | **TALAIRACH** |
|---|---|---|
| **Spatial resolution** | ROI of irregular shape | 1-4$^3$mm$^3$ cubic voxel |
| **Precision** | Usually decreases noise | Introduces additional noise |
| **Efficiency** | Deals with smaller data sets | Deals with larger data sets |
| **Complexity** | Very easy | Extraordinarily complex |
| **Background knowledge** | No | Yes |

A second difficulty that arises when training multiple-subject classifiers is that the intensity of fMRI response to a particular stimulus is usually different across subjects. We employ the same normalization method described in Section 5.1 to linearly re-scale the data from different subjects into the same range to partially address this issue. While there are many inter-subject differences that cannot be addressed by this simple linear transformation, we have found this normalization to be useful in the Sentence versus Picture study, but not in the Syntactic Ambiguity study. As a result, we will not report our results with normalization in Syntactic Ambiguity study.

In order to show whether our results are significantly better than random guessing, we compute the 95% confidence intervals[2] of the accuracies we got from training cross-subject classifiers. Assume that our prediction accuracies are i.i.d. Bernoulli($p$) distributed, so the number of observed correct classifications $X$ will follow a Binomial($n, p$) distribution, where $n$ is the number of test examples. Let $\hat{p} = \frac{X}{n}$ be the observed accuracy. Thus,

$$\begin{aligned} se &= \sqrt{\text{Var}\,(\hat{p})} = \sqrt{\text{Var}\left(\frac{X}{n}\right)} = \sqrt{\frac{p(1-p)}{n}} \\ \widehat{se} &= \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\ CI(\hat{p}) &= (\hat{p} - 1.96\widehat{se}, \hat{p} + 1.96\widehat{se}) \end{aligned}$$

The lowest accuracies that could be regarded as significant depend on specific studies because there are different numbers of examples in different studies, as shown in Table 4.

[2] In fact, the confidence intervals we compute here are the upper bounds of the true confidence intervals under cross validation, which could be shown easily using Lagrangian method.

Table 4: The lowest accuracies that are significantly better than random guessing

| | SP | PS | SP+PS | SYNTACTIC AMBIGUITY |
|---|---|---|---|---|
| **# of Examples** | 520 | 520 | 1040 | 100 |
| **Lowest Accuracy** | 54.3% | 54.3% | 53.1% | 59.7% |

### 6.1 ROI Mapping

The results of multiple subject classifiers using the ROI mapping method in the two studies are presented in Table 5 and Table 6. For the Sentence versus Picture study, the results shown in Table 5 are highly significant compared to the 50% accuracy expected of random guessing, indicating that it is indeed possible to train a classifier to capture significant subject-independent regularities in brain activity that are sufficiently strong to detect single-interval cognitive states in human subjects outside the training set.

Table 5: The accuracies of cross-subject classifiers in the Sentence versus Picture study

| METHOD | CLASSIFIER | SP | PS | SP+PS |
|---|---|---|---|---|
| Average | GNB | 81.7% (88.8%) | 68.1% (82.3%) | 71.5% (74.3%) |
| Average | SVM | 79.8% (86.5%) | 70.4% (77.1%) | 70.8% (75.3%) |
| Average | 1NN | 79.4% (84.8%) | 64.6% (73.8%) | 65.6% (63.7%) |
| Average | 3NN | 82.9% (86.5%) | 65.8% (75.8%) | 67.5% (67.3%) |
| Average | 5NN | 85.4% (88.7%) | 67.9% (78.7%) | 68.7% (68.3%) |
| ActiveAvg(20) | GNB | 84.2% (92.5%) | 69.6% (87.3%) | 68.4% (72.8%) |
| ActiveAvg(20) | 1NN | 84.8% (91.5%) | 64.2% (83.8%) | 67.4% (66.0%) |
| ActiveAvg(20) | 3NN | 86.9% (93.1%) | 66.5% (86.2%) | 69.0% (71.5%) |
| ActiveAvg(20) | 5NN | 87.1% (93.8%) | 68.8% (87.5%) | 70.6% (72.0%) |

An obvious trend in Table 5 is that the accuracies with normalization in SP or PS are much better than the full SP+PS data set. The explanation for this improvement is that the classification task is easier here than when using the full data – in the full data examples come from a greater diversity of temporal contexts, and the effects of these different contexts can remain apparent for several seconds due to the temporally delayed BOLD response. Also, the accuracies in SP are better than in PS, which is consistent with the trend in Table 1. We conjecture that this phenomenon occurs because making a mental picture of a sentence's semantic content is easier than extracting and mentally rehearsing a semantic sentence from a picture, therefore making it harder to separate the two cognitive states in PS. Another conjecture is that the cognitive state of the subjects and the way they respond to a new sentence may be influenced when they have just seen a picture which they expect to compare to the upcoming sentence [7].

Comparing Table 1 with Table 5, we can find another interesting apparent trend that the accuracy on the left out subject for the multiple subject classifiers is often very close to the average accuracy of the single subject classifiers, and in several cases it is statistically significantly better than the corresponding single subject classifiers. This result is surprisingly positive, and it means that the accuracy of this multiple subject classifier, when tested on new subjects outside the training set, is comparable to the average accuracy achieved when training on data from the test subject itself. Presumably this better performance by the multiple subject classifier can be explained by the fact that it is trained using an order of magnitude more training examples, from twelve subjects rather than one. The decrease in the sparsity of the training set after combining data from different subjects greatly com-

pensates for the variability among subjects. Solving the imbalance between the number of features and examples plays a crucial rule in training classifiers in the field of fMRI data analysis.

Table 6: The accuracies of cross-subject classifiers in the Syntactic Ambiguity study

| METHOD | CLASSIFIER | ACCURACY |
|---|---|---|
| Average | GNB | 58% |
| Average | SVM | 54% |
| Average | 1NN | 56% |
| Average | 3NN | 57% |
| Average | 5NN | 58% |
| ActiveAvg(20) | GNB | 61% |
| ActiveAvg(20) | 1NN | 57% |
| ActiveAvg(20) | 3NN | 52% |
| ActiveAvg(20) | 5NN | 57% |

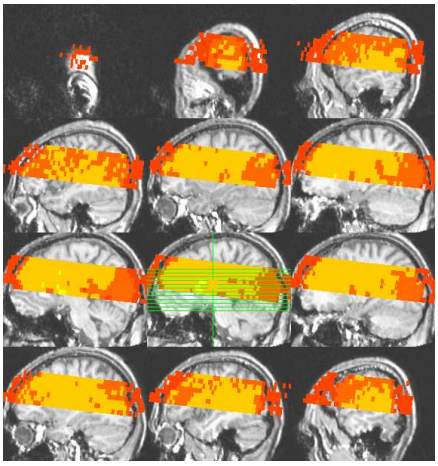
In the Syntactic Ambiguity study, although only one accuracy is significantly better than expected from a random classifier, 50%, no accuracy here is lower than 50%. Unlike the Sentence versus Picture study, the number of examples is much less in this study, hence the imbalance between the number of features and examples becomes much worse. For the same reason, even the single subject classifier accuracy (see Table 2) is not comparable to the ones in the Sentence versus Picture study. We also found that these results are quite sensitive to the particular selection of learning method and feature selection. The cognitive states defined in this study might be too subtle to tell them apart easily[3]. Although we cannot draw strong conclusions from the results we got in this study, it provides modest additional support for the feasibility of training multiple subject classifiers using ROI mapping.

### 6.2 Talairach Transformation

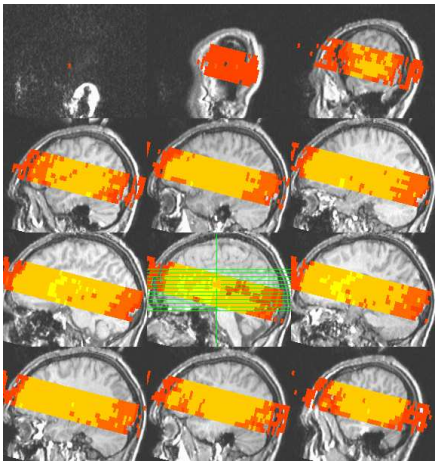
This section will focus on the Syntactic Ambiguity study only[4]. Talairach transformation converts the data from different subjects, through a number of interpolations, into the standard Talairach space which is a box containing the brain. The voxels outside of the brain which should have zero activity are completely useless. We can set some threshold to exclude them out. Furthermore, as shown in Figure 3, focusing on Plot (a) and (b), we can find that the scanning bands in different subjects do not overlap exactly. This means that we cannot directly combine the data from different subjects after Talairach transformation. Otherwise, at some voxels, the activity values are very close the true values in some subjects, but they are zeros in other subjects because these voxels are out of the scanning bands in those subject and actually did not get measured. In response to this discrepancy, we raise the threshold to exclude the voxels out of the intersection of the scanning bands although they are in the brain. In Plot (c), the curve shows that we begin to go into the intersection of the scanning bands of all subjects when the threshold is raised to around 500. Using threshold 500, the selected voxels are displayed in Plot (d). The voxels left have an activity value above 500 in all subjects. In total, 5449 voxels are finally selected each having a volume 4mm$\times$4mm$\times$4mm. The results in this section are based upon the data set consisting of the 5449 voxels.

[3]This experiment is equivalent to distinguishing affirmative and negated trials in the Sentence versus Picture study where we were unable to separate them using classifiers [9].

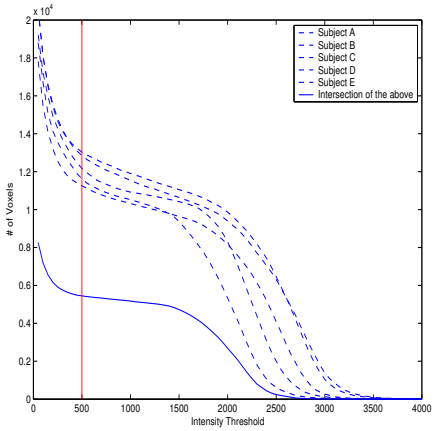
[4]In the Sentence versus Picture study, we have gotten satisfactory results using the ROI mapping method, and we have some technical difficulties in doing Talairach transformation for this study (see [4] for details).



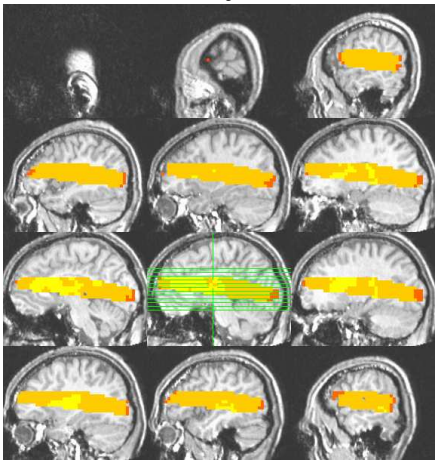
(a) Subject D



(b) Subject E




(c) # of Intersected Voxels



(d) Intersected Voxels

Figure 3: Threshold selection after Talairach transformation. Plot (a) and (b) show the functional data of two subjects (Syntactic Ambiguity study) in Talairach space in sagittal view. The scanning bands differ significantly in these two subjects. Plot (c) is the curve of the number of voxels left as the intensity threshold changes. Plot (d) provides the intersection of the selected voxels in all subjects using threshold 500 in Talairach space.

We trained multiple subject classifiers on the whole data set (5449 voxels) and the best accuracy produced was 59% using SVM. Note that the severe imbalance between number of features and examples ($5449 \times 16$ versus 100). It is natural to employ the feature selection methods used in training single subject classifiers to reduce the number of features because we treat the data as if they were from the same subject. The only difference is that in this case we have no ROI information to use in some abstraction operation[5]. Here, we mainly explored $Active(n)$ (described in Section 4) with all possible even numbers smaller than 200. The results, shown in Table 7, are all significantly better than random guessing. In contrast to Table 2, the accuracy of this multiple subject classifier again is comparable to the average accuracy of single subject classifier. Although the results in Table 7 look better

[5]We tried to cluster the voxels according to their geographic locations, but we did not get any obvious clusters.

Table 7: The best accuracies of cross-subject classifiers using $Active(n)$. The third column denotes the number of voxels we used to get such an accuracy.

| CLASSIFIER | BEST ACCURACY | # OF VOXELS |
|---|---|---|
| GNB | 63% | 96,98,100,114 |
| SVM | 67% | 138 |
| 1NN | 60% | 124 |
| 3NN | 60% | 192,194,196 |
| 5NN | 62% | 24 |

than the ones in Table 6, we cannot use them to draw any conclusion about which method is better. Table 7 presents the best accuracies after we explored many possible settings, but Table 6 only provided the results under some particular setting.

## 7 Conclusions

The primary goal for this research was to determine whether it is feasible to use machine learning methods to decode mental states across multiple subjects. The successful results reported above for two studies indicate that this is indeed feasible in a variety of interesting cases. However, it is important to note that while our empirical results demonstrate the ability to successfully distinguish among a predefined set of states occurring at specific times while the subject performs specific tasks, they do not yet demonstrate that trained classifiers can reliably detect cognitive states occurring at arbitrary times while the subject performs arbitrary tasks. While our current results may already be of use in cognitive science research, we intend to pursue this more general goal in future work.

Two methods were explored to train cross-subject classifier based upon fMRI data. The ROI mapping method abstracts the fMRI data by using the mean fMRI activity in each of several anatomically defined ROIs to map different brains in terms of ROIs. Talairach transformation provides another way to match different brains at finer spatial level. Using these approaches, it was possible to train classifiers to distinguish, e.g., whether the subject was viewing a picture or a sentence describing a picture, and to apply these successfully to subjects outside the training set. In most cases, the classification accuracy for subjects outside the training set equalled or exceeded the accuracy achieved by training on data from just the single subject.

### Acknowledgments

Many thanks go to Prof. Marcel Just for valuable discussion and suggestions. We would also like to thank Francisco Pereira and Radu S. Niculescu for providing much code to run our experiments. I am grateful to Vladimir Cherkassky, Joel Welling, Erika Laing and Timothy Keller for their instruction on many techniques related to coregistration and Talairach transformation.

### References

[1] Blankertz, B., Curio, G., & Mller, K. R., Classifying Single Trial EEG: Towards Brain Computer Interfacing, *Advances in Neural Information Processing Systems (NIPS 2001)*, vol. 14, 157-164, MIT Press, 2002.

[2] Burges C., A Tutorial on Support Vector Machines for Pattern Recognition, *Journal of data Mining and Knowledge Discovery*, 2(2),121-167, 1998.

[3] Haxby, J., Gobbini, M., Furey, M., Ishai, A., Schouten, J., & Pietrini, P., Distributed and over-

lapping representations of faces and objects in ventral temporal cortex, *Science*, 293, 2425-2430, 2001.

[4] Keller, T., Just, M., & Stenger, V., Reading span and the time-course of cortical activation in sentence-picture verification,*Annual Convention of the Psychonomic Society*, Orlando, FL, 2001.

[5] Mason, R., Just, M., Keller, T., & Carpenter, P., Ambiguity in the brain: What brain imaging reveals about the processing of syntactically ambiguous sentences, *in press*, 2003.

[6] Mitchell, T., *Machine Learning*, McGraw-Hill, 1997

[7] Mitchell, T., Hutchinson, R., Just, M., Newman, S., Niculescu, R., Pereira, F., & Wang, X., Machine Learning of fMRI Virtual Sensors of Cognitive States, *The 16th Annual Conference on Neural Information Processing Systems*, Computational Neuroimaging: Foundations, Concepts & Methods Workshop, 2002

[8] Mitchell, T., Hutchinson, R., Just, M., Niculescu, R., Pereira, F., & Wang, X., Classifying Instantaneous Cognitive States from fMRI data, *The American Medical Informatics Association 2003 Annual Symposium*, submitted, 2003

[9] Mitchell, T., Hutchinson, R., Niculescu, R., Pereira, F., Wang, X., Just, M., & Newman, S., Learning to Decode Cognitive States from Brain Images, *Machine Learning: Special Issue on Data Mining Lessons Learned*, submitted, 2003

[10] Talairach, J., & Tournoux, P., *Co-planar Stereotaxic Atlas of the Human Brain*, Thieme, New York, 1988.

[11] Wagner, A., Schacter, D., Rotte, M., Koutstaal, W., Maril, A., Dale, A., Rosen, B., & Buckner, R., Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity, *Science*, 281, 1188-1191, 1998.